Network Quality-of-Service Demonstration at Supercomputing 97

Jude A. George[1]

NAS Technical Report NAS-98-006 June 98

jude@nas.nasa.gov
FSC End2End, Inc.
NAS Wide Area Network Group
NASA Ames Research Center
Mail Stop 258-5
Moffett Field, CA 94035-1000

Abstract

Traditional best-effort data networks do not meet the requirements of computing environments which support multiple users of high-reliability, high-availability, and high-capacity applications. To approach this problem, a data network must be designed to provide each user real-time and scheduled access to resources without interference from other users' applications. An ATM network supporting real-time video applications utilizing full OC-12 bandwidth was built and demonstrated at the Supercomputing 97 trade show in San Jose, California. Described herein are the experiments conducted on this network and the lessons learned.

---

1. FSC End2End Inc., for MRJ, Inc., NASA Contract NAS2-14303, Moffett Field, CA 94035-1000

This page intentionally left blank.

1.0 Introduction and Purpose

NASA's high-performance networks have traditionally been used by the scientific computing community for file transfer between supercomputers, mass storage, and workstations. The obstacle for users to do real-time work across a network has been the lack of an enabling technology as well as an architecture that allows the end-user to be guaranteed a desired level of network service. For a real-time application, network delay, jitter, and available bandwidth must not impose limitations on what can be accomplished. The user should be able to use a remote resource as if it were on the local workstation.

This report describes a network quality-of-service experiment at the Supercomputing '97 conference in November 1997 in San Jose, California. For the purposes of our demonstration, we use the term quality-of-service to mean the delivery of a guaranteed service level, rather than a "best-effort" scheme such as RSVP. ATM was chosen as the technology to investigate network quality-of-service because of it's bandwidth and delay characteristics, it's maturity of implementation, and the availability of hardware.

To demonstrate the bandwidth and delay properties of ATM networks operating at the edge of their performance envelope, we used two applications on the Supercomputing '97 show floor. The first is a digital video editing environment which carries video, audio, and control information across a single ATM PVC. The second is a videoconferencing and remote monitoring application which carries video, audio, and control across three separate PVCs.

We will begin by describing the applications that were used to test the capability of the network to support real-time work, and the software used to reserve bandwidth on the network. We will then describe the network architecture created for the tests, and then the tests themselves, followed by the results of the tests. Documents describing the flow analysis of the network [1] and the network control software [2] are available from their respective authors.

2.0 The Applications

2.1 SMPTE 259M Digital Video

The creation of "contribution level video", the highest quality used in video production, requires uncompressed video streams to be transported between editing and storage systems. The SMPTE 259M standard defines a digital video data stream which uses 270Mb/s of

network payload bandwidth [3]. The application consists of sending a video stream with embedded sound and control data [4], [5] from an Ampex digital tape deck ("Tape") to an Avid Media Composer video editing station ("Editor"), and back from the editing station to the digital tape. This application simulates the use of an ATM network for remote editing between a production studio and a user situated at an off-site facility.

The Editor and Tape video streams are encapsulated into ATM AAL5 frames by the Tektronix Cascadable Video Interface, or CVI. Because of this, the Editor and the Tape themselves are not considered to be part of the ATM network architecture.

This application effectively stresses the network's capability to provide a consistent level of bandwidth with very narrow tolerances for delay and jitter.

2.2 NTSC Video

For videoconferencing and remote collaboration, the quality of NTSC video is usually deemed to be sufficient. The application consists of sending video from the twelve-foot wind tunnel at the NASA Ames Research Center through an ATM network to a display situated on the show floor. This application simulates the use of the DARWIN network for remote wind tunnel test analysis by researchers at remote DARWIN customer sites, such as airframe manufacturers.

2.3 IP Connectivity

IP (Internet Protocol) connectivity allows for login and file transfer sessions between workstations. Our goal with this experiment was to determine if the experimental network could support general IP connectivity while maintaining the higher levels of service for video. To test IP connectivity we used the HNMS network monitoring system and the ForeView network monitoring system.

3.0 Network Architecture

3.1 Show-Floor Network

The experimental ATM network built for these demonstrations, known as SCinetEx or X-net, consisted of multiple ATM switches interconnecting the video devices to the show-floor production network (SCinet) and to the wide-area network. Three show-floor switches were set up for the SMPTE 259M video experiments. An additional switch,

which also carried traffic for the NASA booth, was used to display NTSC video in both the NASA booth and the X-net booth. All of these switches, plus one switch in the MRJ, Inc. booth, were also used in the IP network.

3.2 Wide-Area Connectivity

The wide-area connectivity was provided by the National Transparent Optical Network, or NTON, at the show floor and at NASA Ames. The NTON is a WDM (wavelength division multiplexed) OC-48 backbone surrounding the SF Bay Area providing connectivity to participating research, corporate, and education sites. Two OC-12 drops were provided into the SCinet at the show floor, and one OC-3 was provided into the NREN (National Research and Education Network) ATM switch at NASA Ames. The OC-12 drops were used to support our SMPTE 259M application, and the OC-3 drop was used to bring NTSC video onto the show floor from the DARWIN network at NASA Ames. An additional OC-12 was provided at NASA Ames for a separate application.

3.3 Network Control Software

The GRAMS (Global Resource Accounting and Management Software) package, which was developed for delivering guaranteed network service levels in this type of environment, was used to reserve, set-up, and tear-down network connections. GRAMS provided a web interface accessible from the SGI Indigo2 workstation used for network control and monitoring, as well as from a PC used by the end-users, and a workstation at a remote booth on the show floor. The web interface allowed the users to reserve network bandwidth and end devices, and to query and delete existing reservations. Confirmation of reservations and billing information was emailed to the users. Once a reservation was created, the software used predetermined path information to determine which ATM switches needed to be configured to create the required network path for the user, and configured PVCs on the switches via the Fore AMI remote configuration interface. At the expiration of the reservation, the switches were again contacted and the PVCs were removed.

4.0 Experiments

4.1 SMPTE 259M Digital Video over ATM

A video editing environment using the Avid Media Composer editing station and the Ampex digital tape recorder was built within the X-net

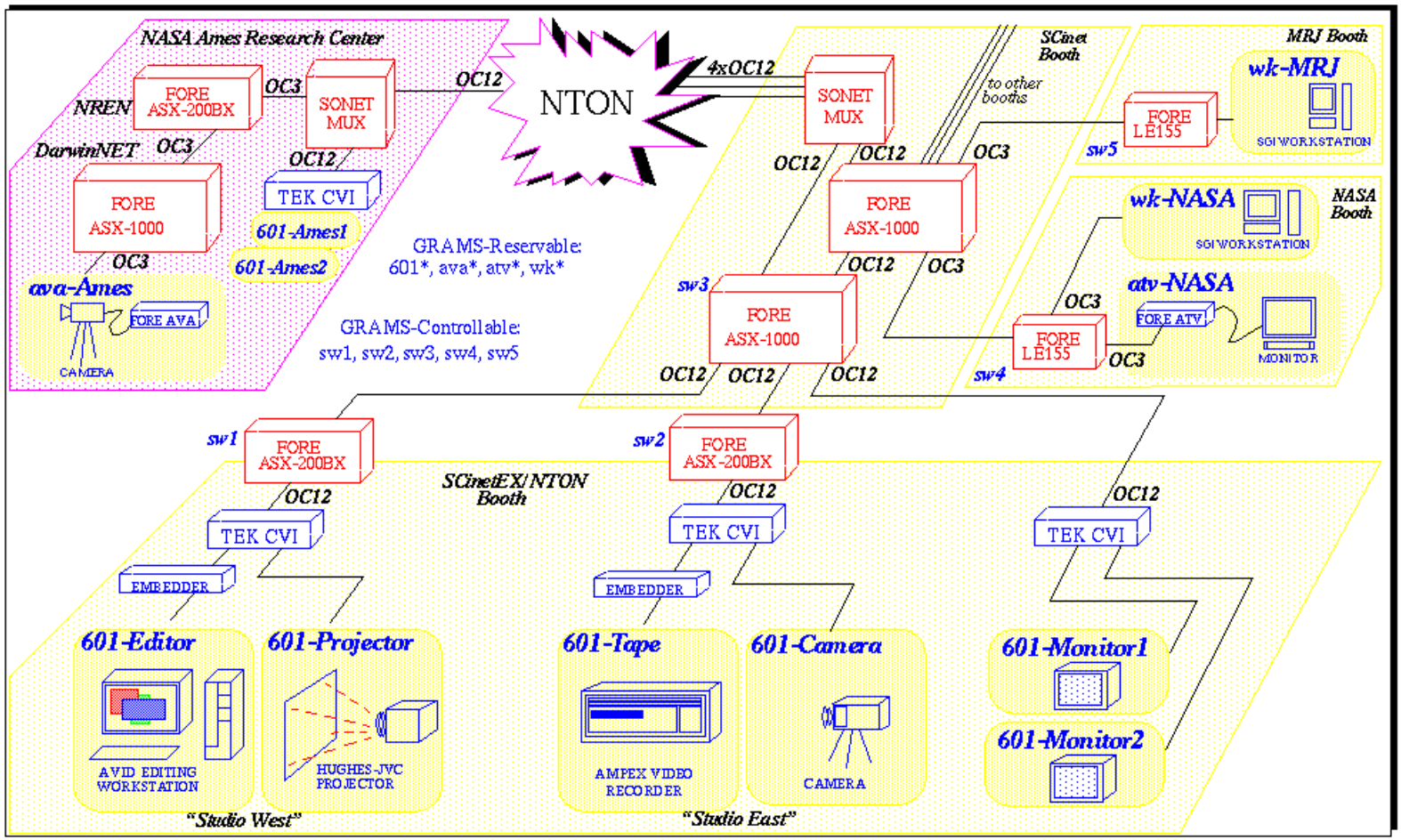


FIGURE 1. Experimental ATM Network

booth. As shown in Figure 1, the Editor and Tape were each connected to the ATM network using a Tektronix CVI board, or Cascadable Video Interface. Each CVI board is designed to carry two 259M video streams on a single OC-12 ATM/SONET link using AAL5. Figure 1 also shows the projector and camera as being ATM-attached in the network; for the duration of our experiments, however, the projector was connected directly to the editor, and the camera was not used. All interfaces through which the video traffic passed were OC-12 interfaces; single-mode interfaces and fiber were used between the switches and the CVIs, and multi-mode interfaces and fiber were used between the switches.

For evaluating the quality of the video signal, oscilloscopes (Tektronix WFM601M Video Waveform Monitors) were connected in series with the video streams as they entered and exited the CVIs. These allowed for the measurement of jitter. We did not have any capability to directly measure ATM cell loss.

The CVIs support error detection via a 32-bit CRC calculation on the data and by comparing the number of cells received with the number of cells that are supposed to be in the AAL5 packet. However, there is no provision for error correction as individual cells do not have sequence numbers, and there is no way to tell which cell(s) may have been dropped. If any cells have been dropped, the CVI resets the cell counter so that synchronization is not lost.

Refer to the switches labeled sw1, sw2, and sw3. Sw1 and sw2 are Fore ASX-200BX switches, each with one switching fabric. Sw3 is a Fore ASX-1000 with two switching fabrics. Each switching fabric in an ASX-1000 can be considered to be a fully independent ATM switch with it's own CPU and software configuration; the two fabrics are joined by a 10GB/s backplane.

By configuring the switches labeled sw1, sw2, and sw3, three distinct network paths were created at different times to conduct the video experiments.

The Fore ATM switches support the use of cell-rate policing with what are known as "UPC contracts". We used CBR (constant bit rate) UPC contracts to set aside a specific amount of bandwidth for each reservation.

4.1.1Two Switches

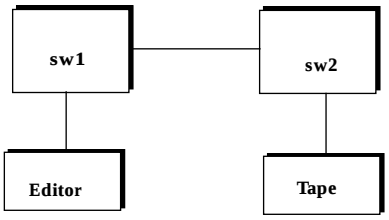
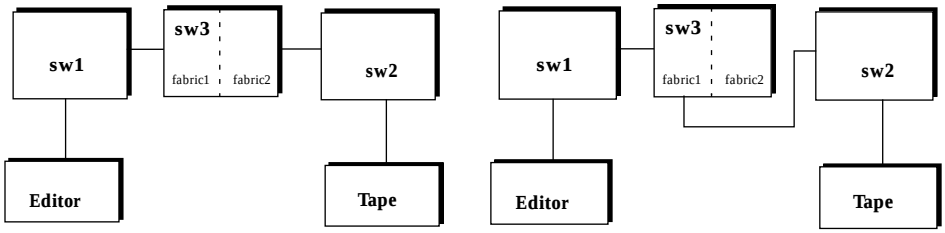


FIGURE 2. Two Switches

The first network configuration in our testing (Figure 2) used a bypass link between sw1 and sw2, such that cells traveled from the Editor, to sw1, directly to sw2, and to the Tape. Sw3 was thus bypassed.

4.1.2Three Switches



(3a) dual fabric test (3b) single fabric test

FIGURE 3. Three Switches

The second network (Figure 3) configuration required traffic generated by the Editor to enter sw1, leave the X-net booth to go to sw3 in the SCinet booth, return to the X-net booth to enter sw2, and proceed to the Tape. In turn, traffic from the Tape to the Editor would take the reverse path. Our first test using this physical layout (Figure 3a) included configuring sw3 to have a video stream enter and exit the switch through separate switching fabrics, requiring the cells to traverse the 10GB/s backplane of the switch which connects the fabrics. Our second test (Figure 3b) involved configuring the switch (and moved the appropriate interfaces) to have the cells enter and exit sw3 through the same fabric.

4.1.3Three Switches and NTON

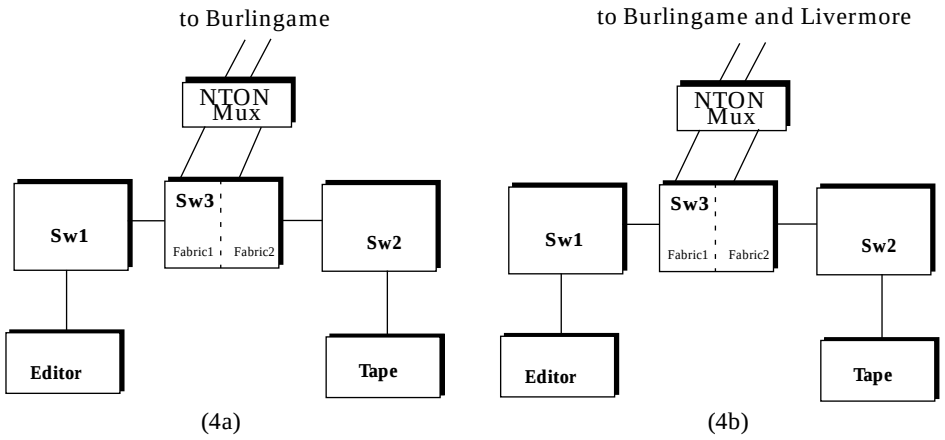


FIGURE 4. Three Switches and NTON

The third network path (Figure 4) required traffic to proceed from the Editor to sw1, over to sw3, out over the NTON, back from the NTON into sw3, and into the Recorder. Again, the path supports traffic in the reverse direction as well.

In the NTON setup, two separate wide-area network paths were used. In the first path (Figure 4a), traffic went from the show floor out to Burlingame and back. In the second path (Figure 4b), traffic went from the show floor to Burlingame, to Livermore, back to Burlingame, and back to the show floor.

4.2 NTSC Video over ATM

NTSC video was transported over the ATM network using Fore AVA and ATV devices. The AVA is an NTSC-to-ATM encoder, and the ATV is an ATM-to-NTSC decoder. These codecs require three VCs: one for video, one for audio, and one for control information. These VCs may be either SVCs or PVCs. For SVCs to be used, it is required that the AVA and ATV be configured and controlled using Fore's own SVA software. The hardware supports the use of PVCs if the PVCs are set up by hand. In our case, the PVCs were configured by hand to establish a path from an AVA in the 12-foot wind tunnel at NASA Ames, to the DARWIN hub switch, to the Ames NREN switch, to the Ames NTON mux, to the show floor's NTON mux, to sw3, to the show floor NASA switch, and finally to the ATV in the NASA booth. This ATV was temporarily relocated to the ScinetEx booth by extending the ATM link from the NASA booth switch to the NASA ACTS switch in the ScinetEx booth.

4.3 IP over ATM

IP over ATM was used in the test network primarily for monitoring the switches. A single Logical IP Subnet (LIS) was built, with IP addresses for the following devices: sw1, sw2, sw3, sw4, sw5, wk-MRJ, and the GRAMS workstation. Although sw3 had two switching fabrics, only one of them could be assigned an IP address at any given time.

5.0 Results

5.1 SMPTE 259M Digital Video over ATM

5.1.1Two Switches

With two switches, there were no visible problems using one stream in each direction.

The network was able to support two simultaneous video streams in each direction with some momentary loss of quality; there were slight glitches lasting less than one second. The problem was probably the overrunning of buffer space in the switches, specifically in the single-mode OC-12 cards, which were an old revision. Buffer utilization was monitored in the CVIs, and was not found to be a cause of any data loss.

5.1.2Three Switches

Our first three-switch configuration had the video stream passing through both fabrics of the switch. Using this configuration, we discovered that we could not create PVCs between the two switching fabrics of sw3 in the expected fashion; that is, we could not create separate PVCs into and out of the switch's backplane, with VC identifiers of our choice. A telephone discussion with one of Fore's engineers verified that it was necessary to have Fore's SPANS SVC protocol configured when configuring PVCs across the backplane of a multi-fabric switch. Using the SPANS protocol, we were allowed to choose the VC id's on the external interfaces of the switch, but had to rely on SPANS to choose the VC identifiers on the backplane itself. In effect, SVCs are created within the switch.

Unfortunately, using SPANS to create the PVCs results in an information gap for any outside software package that is expected to control the use of PVCs in the network, since the VCI of each PVC crossing the backplane is not known until *after* it is created. Fortunately, we were able to reconfigure GRAMS to use this scheme without having to modify any code. In our second configuration with sw3, this problem was bypassed altogether by ensuring that the video stream did not cross

the switch's backplane between the two fabrics.

In the three-switch network, using one video stream, there were periodic impairments that showed on the oscilloscopes, but not on the screen. The frequency of the dropouts was twice as often when we went from one stream to two.

With the addition of a second video stream in the same direction on the OC-12, the video signal would show a complete loss of video for approximately one second at a time. A complete loss of signal or synchronization is evident because in the absence of an input signal, the VPG embedders insert a green screen, which is what was shown on the output. The problem was probably due to SONET slip or a buffer overrun in the ATM single-mode cards, which were of an old revision. There were no buffer overruns in the CVIs.

If we had synchronized all switches to a single clock, we may not have experienced this problem. However, we did not learn of a way to accomplish during the show, if indeed it was possible with our hardware.

5.1.3Three Switches and NTON

With three switches plus the NTON connection, the results were the same as those for just three switches in the show-floor network. We were able to pass a single stream of video through the show-floor switches and the NTON without any video dropouts, but the addition of a second stream in the same direction resulted in momentary signal loss.

5.1.4Jitter Measurement

Jitter on the SMPTE 259M video signal was measured using the Video Waveform Monitors and was found to increase with the addition of the second video signal. The measurement was made in terms of the Unit Interval (UI), or clock period. The increased jitter did not visibly affect the video image quality.

TABLE 1. Video Jitter, via Loopback through ASX-200BX

Number of SMPTE 259M Video Streams	Peak-to-Peak	With 100kHz filter	With 1kHz filter
2 Channels of Video, Both Directions; Measurement on Channel 2	~3 UI	0.3 UI	0.9 UI
1 Channel of Video, Both Directions; Measurement on Channel 2	~1.5 UI	0.2 UI	0.47 UI

TABLE 2. Video Jitter, via Loopback through ASX-200BX and ASX-1000

Number of SMPTE 259M Video Streams	Peak-to-Peak Jitter	With 100kHz filter	With 1kHz filter
2 Channels of Video, Both Directions; Measurement on Channel 2	~3 UI	0.2 UI	0.9 UI
2 Channels of Video, Both Directions; Measurement on Channel 1	~3-4 UI	0.2 UI	0.9 UI
1 Channel of Video, Both Directions; Measurement on Channel 2 (signal from Digital Tape)	~1.3 UI	0.2 UI	0.9 UI
1 Channel of Video, Both Directions; Measurement on Channel 1 (signal from color bar generator)	~2.5 UI	0.2 UI	0.5 UI

5.1.5Bandwidth Measurement

By pre-defining CBR UPC contracts on the switches, and successively increasing the CBR cell rate from a rate known to be less than what is required, we measured the ATM bandwidth required for a single video stream to be 298Mb/s. This is almost exactly what is expected given a 270Mbps data rate with a 5 byte overhead per 53 byte ATM cell.

TABLE 3. Bandwidth Measurement through ASX-200BX

Bidirectional CBR (Constant Bit Rate) Bandwidth Allocated on each Channel	Result
271 Mbps	no video on either channel
299 Mbps	both channels have excellent video
298 Mbps	both channels have excellent video
290 Mbps	no video on either channel
295 Mbps	no video on either channel
297 Mbps	no video on either channel
297.5 Mbps	no video on either channel

5.2 NTSC Video over ATM

During our show-floor demonstrations, PVCs carrying the video and control information for the NTSC video stream were set up and removed manually, as we had not configured GRAMS to support the NTSC experiment. However, prior demonstrations in the NAS data communications lab verified that the NTSC video worked as expected when configured via GRAMS; these demonstrations carried audio as well as the video and control channels.

The AVA/ATV devices we used had the ability to be remotely controlled via a software package. However, this functionality required that SVCs be used for the video, audio, and control connections, rather than PVCs. Since our GRAMS software only had the ability to manipulate PVCs, we would not have been able to use the remote control features in any case, which would have allowed us to switch between multiple video inputs on the AVA. The video input was manually switched between a source showing a DARWIN wind tunnel test, and a live space shuttle launch.

No ill effects were observed in the show-floor NTSC video.

5.3 IP over ATM

We encountered problems when setting up a Logical IP Subnet (LIS) across the backplane of the ASX-1000 using Classical IP. When the LIS was configured using SVCs across both switching fabrics of sw3, a problem would develop where certain end stations would become disjoint from the LIS and were not able to ping the other end stations, thus violating the RFC 1577 specification for IP over ATM.

As a result, we used an out-of-band (Ethernet) connection to each host and switch for network management purposes. We later discovered that using SPANS instead of Classical IP resulted in the LIS remaining intact.

5.4 Connection Management

The GRAMS software effectively managed the bandwidth of the network by allocating CBR PVCs between all of the ATM-attached devices. Although we only used the Editor and Tape as edge devices, GRAMS was configured to allocate bandwidth for the Camera and Projector as well. Set-up and tear-down time for the connections was approximately one second in all cases, end-to-end; the multiple devices for each connection were configured by the software in parallel. The software did not allow oversubscription of any network resource, or simultaneous allocation of any of the edge devices to multiple users.

6.0 Conclusions

We have shown that ATM technology can easily support the capacity requirements of multiple high-bandwidth users, especially with a scheduling and reservation system in place. Our choice for a high-bandwidth application -- SMPTE 259M video -- was a good one, as it was shown to be very sensitive to even momentary reductions in network quality of service.

From the users' perspective, guaranteed service levels were provided by the GRAMS control software by allowing the users to reserve in advance what resources they need, and having those resources be available when it came time for their applications to be used. Other users were prevented from having access to bandwidth allocated for another user. Of course, the ability to guarantee service levels in this fashion requires a highly controlled environment in which there are no rogue users, and in which all devices support the capability to reserve bandwidth and preserve the delay characteristics of traffic passing through them.

The devices we tested show need for improvement in the areas of buffer utilization and SONET timing. It is possible that if we had obtained current revisions of the single-mode OC-12 ATM cards, and had the ability to configure the switches to synchronize their clocks, we would have eliminated these problems.

Our IP connectivity problems using ATM SVCs showed that having out-of-band connectivity is important in a scenario where each switch (and potentially each host) must be independently reachable via IP in order to configure it, either manually or via a software package. An alternative would be to create a fully-meshed PVC network for IP connectivity to each switch, but this is prohibitive when using a large number of switches. Ideally, we would be able to rely on IP over SVCs, and use out-of-band connections as a backup measure.

The impact of the lack of support in our ATM switches for the creation of PVCs between switching fabrics suggests that automated configuration software (such as GRAMS) should allow the retrieval of configuration information when the switches themselves allocate connection identifiers such as VPIs and VCIs. The restricted feature set of the AVA/ATV NTSC video devices when using PVCs supports this conclusion; generally, devices will have more functionality when SVCs are used.

By having some of the resource decision-making being relegated to the hardware devices, determinism is not necessarily lost since the controlling software still has full knowledge of the state of the hardware configuration after the connection has been created. This type of dissociated control will be the first step toward manipulating SVCs with GRAMS or other software packages.

7.0 Hardware and Software

The following equipment and software were used to conduct the experiments described in this paper. Note, however, that the extended network shown in the diagram had additional equipment not used in our experiments.

7.1 SCinetEX Booth:

1. Fore Systems ASX-200BX ATM switch:
   - Fore Systems NM-1/622SMSCC 1-port single-mode OC-12
   - Fore Systems NM-1/622MMSCC 1-port multi-mode OC-12
2. Fore Systems ASX-200BX ATM switch:
   - Fore Systems NM-1/622SMSCC 1-port single-mode OC-12
   - Fore Systems NM-1/622MMSCC 1-port multi-mode OC-12
3. Fore Systems ASX-1000 ATM switch w/ 2 switching fabrics:
   - Fore Systems NM-1/622MMSCC 1-port multi-mode OC-12 (x5)
4. Tektronix VME chassis (x2):[1]
   - Tektronix 259M Cascadable Video Interface
   - Tektronix OC-12 SONET/ATM Network Interface
5. Video Products Group 259M Embedder (x2)
6. Ampex DCT
7. Hughes/JVC Projector
8. Tektronix WFM601M Video Waveform Monitors
9. Silicon Graphics Indigo 2 w/ 64MB RAM, 2GB disk
   - NASA/FSC End2End GRAMS Software
   - Fore Systems OC-3 Network Interface
   - Fore Systems ForeView Software

7.2 NASA Booth:

1. Fore Systems ASX-200BX ATM switch:
   - Fore Systems NM-2/155MMSRC 4-port multi-mode OC-3
   - Fore Systems NM-1/622MMSCC 1-port multi-mode OC-12
2. Fore Systems ATV-300/MMOC3 NTSC Video D/A Converter
3. Fore Systems ASX-200 ATM switch:
   - Fore Systems NM-2/155MMSRC 4-port multi-mode OC-3

---

1. One of the chassis contained an additional set of interfaces (CVI and SONET) to support another experiment.

7.3 NASA Ames Research Center:

1. Fore Systems ASX-1000 ATM switch w/ 1 switching fabric:
   - Fore Systems NM-2/155SMSRC 4-port single-mode OC-3
2. Fore Systems AVA-300/MMOC3 NTSC Video A/D Converter

7.4 Service Network Equipment:

The following network devices served as a pass-through for our experimental traffic. PVCs were requested from the administrators of these networks to support the experimental traffic.

1. NREN switch at NASA Ames Research Center
2. NTON SONET multiplexer at NASA Ames Research Center
3. NTON SONET multiplexer at Supercomputing '97

8.0 Acknowledgments

Many thanks go to F. Ronald Bailey, Bruce Blaylock, Chuck Garsha, Gary Goncher, Arshad Khan, Kevin Lahey, Jim McCabe, and Keith Nesson for reviewing this report and/or making available information from their notes.

9.0 References

[1] “Network Requirements and Flow Analyses for the Network Quality-of-Service Demonstration at SC ‘97”, James D. McCabe, NAS Facility, NASA Ames Research Center, 1997.

[2] “SC97 Global Resource Accounting and Management Design Document”, Jude A. George, NAS Facility, NASA Ames Research Center, Nov. 1997.

[3] ANSI/SMPTE 259M, for Television -- “10-bit 4:2:2 Component and NTSC Composite Digital Signals -- Serial Digital Interface”, SMPTE, White Plains, NY, Sept. 1997.

[4] ANSI/SMPTE 291M, for Television -- “Ancillary Data Packet and Space Formatting”, SMPTE, White Plains, NY, Jan. 1996.

[5] ANSI/SMPTE 12M, for Television, Audio, and Film -- “Time and Control Code”, SMPTE, White Plains, NY, Sept. 1995.

NAS	**NAS TECHNICAL REPORT**
	Title:
	Author(s):
	Reviewers: "I have carefully and thoroughly reviewed this technical report. I have worked with the author(s) to ensure clarity of presentation and technical accuracy. I take personal responsibility for the quality of this document."
Two reviewers must sign.	Signed: _____ Name: _____ Signed: _____ Name: _____
After approval, assign NAS Report number.	Branch Chief: Approved: _____
Date:	NAS Report Number: